A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications

Year: 2023 | Citations: 593 | Authors: Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, José I. Santamaría, A. Albahri

---

Abstract

Data scarcity is a major challenge when training deep learning (DL) models. DL demands a large amount of data to achieve exceptional performance. Unfortunately, many applications have small or inadequate data to train DL frameworks. Usually, manual labeling is needed to provide labeled data, which typically involves human annotators with a vast background of knowledge. This annotation process is costly, time-consuming, and error-prone. Usually, every DL framework is fed by a significant amount of labeled data to automatically learn representations. Ultimately, a larger amount of data would generate a better DL model and its performance is also application dependent. This issue is the main barrier for many applications dismissing the use of DL. Having sufficient data is the first step toward any successful and trustworthy DL application. This paper presents a holistic survey on state-of-the-art techniques to deal with training DL models to overcome three challenges including small, imbalanced datasets, and lack of generalization. This survey starts by listing the learning techniques. Next, the types of DL architectures are introduced. After that, state-of-the-art solutions to address the issue of lack of training data are listed, such as Transfer Learning (TL), Self-Supervised Learning (SSL), Generative Adversarial Networks (GANs), Model Architecture (MA), Physics-Informed Neural Network (PINN), and Deep Synthetic Minority Oversampling Technique (DeepSMOTE). Then, these solutions were followed by some related tips about data acquisition needed prior to training purposes, as well as recommendations for ensuring the trustworthiness of the training dataset. The survey ends with a list of applications that suffer from data scarcity, several alternatives are proposed in order to generate more data in each application including Electromagnetic Imaging (EMI), Civil Structural Health Monitoring, Medical imaging, Meteorology, Wireless Communications, Fluid Mechanics, Microelectromechanical system, and Cybersecurity. To the best of the authors' knowledge, this is the first review that offers a comprehensive overview on strategies to tackle data scarcity in DL.